Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

From the Buzzing in Turing's Head to Machine Intelligence Contests

Huma Shah[1], **Kevin Warwick** [2]

Abstract. This paper presents an analysis of three major contests for machine intelligence. We conclude that a new era for Turing's test requires a fillip in the guise of a committed sponsor, not unlike DARPA, funders of the successful 2007 Urban Challenge.

1 INTRODUCTION

This paper reviews three current competitions for machine intelligence. All three have featured at least one artificial conversational entity – ACE, as a contestant. ACE are systems that attempt to deceive by *thinking*, described by Turing as a 'buzzing' inside one's head, using text-based human-like linguistic productivity, *li.p*. Based on Turing's imitation game [1], the authors feel a machine's comparison against a human, both simultaneously questioned by an *average interrogator* provides a useful insight into the processes beneath what humans do profusely: talking. Neither the Loebner Prize for Artificial Intelligence [2], the Chatterbox Challenge [3] or BCS's Machine Intelligence Prize [4] discussed here have shown the success of DARPA's 2007 Urban challenge [5], a timed race for autonomous vehicles negotiating traffic following California driving rules. We conclude that what the Turing Test venture requires, to encourage interdisciplinary collaborative teams, is a generous sponsor(s) truly interested in fostering engineering achievements that exhibit extraordinary human progress when faced with an indomitable challenge.

2 TURING TEST - IS THERE ANY POINT?

Turing's textual machine-human comparison test will be discussed ad infinitum, even when a disembodied artificial intellect, which, through its mental capacities, deceives 30% of a panel of human judges - Turing criteria[3], that they are in conversation with another human. Turing circumvented defining intelligence by positing an imaginary experiment in which a machine can respond appropriately to interrogator questions on any topic.

Largely ignored by academia in practical terms, the Turing Test has provided much philosophical fodder for the past six decades, including arguments over how many tests, and whether gender of the human foil is important [6]. Nonetheless, the emergence of Weizenbaum's natural language understanding endeavour [7] through his Eliza system in the mid 1960s has spurred on many an enthusiast to build a conversational partner that surprises its human interlocutor. However, the lucrative digital world of the Internet means some of the best systems are not entered for machine intelligence competitions; hackers exploit the technology for nefarious purposes[4]:

> "New software designed to conduct flirtatious conversations is good enough to fool people into thinking they are chatting with a human ... *CyberLover* software ... to engage people in conversations with the objective of inducing them to reveal information about their identities or to lead them to visit a web site that will deliver malicious content to their computers."

One significant element missed amongst the plethora of Turing Test literature, pointed out by Demchenko and Veselov [8], two thirds of the team behind *Eugene* a ten year-old Ukrainian child-mimicking ACE[5], is the language of conducted tests: English. With its "unexcelled number of idiomatic phrases" and "words with multiple meanings", Demchenko and Veselov draw attention to the manner in which non-native English speaking ACE designers approach English colloquialisms and metonyms. Speakers of Russian, Demchenko and Veselov have a "mechanistic view of grammar construction". They contend that "different languages have varying degrees of success" [8: p.452], and that "modelling and imitation of thinking of people is much easier in some languages than in others" (p. 453). Their biggest problem in designing an English-speaking ACE is culture, and what knowledge is appropriate to inculcate their artificial child chatter with.

A "poorly designed experiment" the Turing Test, is dependent "entirely on the competence of the judge" [9]. In the first Loebner Prize in 1991, "some judges ...rated a human as a machine on the grounds that she produced extended well-written paragraphs of informative text at dictation speed without typing errors" (ibid) – an inhuman feat, according to that particular judge. They also point out the absurdity of *knowing*, or that *believing to know how something works* distorts our view of its intelligence (p.53). Humphrys [10] claims that the Turing Test has been passed, adding indifferently: "so what" (p.256). To the question "Is the Turing test, and passing it, actually *important* for the field of AI?" Humphrys declares, "No" (p. 2.55). He accepts his own creation MGonz is based on trickery and that it has "no AI". Surely then, the goal should be seeking inventive methods for ACE *li.p* expressing apropos emotions reflecting back on utterances during dialogues. Humphrys reminds "to take care that it is not just based on *prejudice*" [10: p. 255]. Indeed, Loebner Prize Turing Test judges have pronounced some ACE

[1] School of Systems Engineering, The Univ. of Reading, RG6 6AY, UK. Email: h.shah@reading.ac.uk
[2] School of Systems Engineering, The Univ. of Reading, RG6 6AY, UK. Email: k.warwick@reading.ac.uk
[3] Turing, 1950: p. 442

[4] *Cyberlover* software developed to steal identities: http://www.itwire.com/content/view/15748/53/ accessed 4.1.10; time 15.32
[5] Eugene Goostman: http://www.princetonai.com/bot/bot.jsp accessed: 4.1.10; time: 16.13

entrants as little more than Eliza duplicates, while wrongly ranking others as human[6].

3 LOEBNER PRIZE 1991 -

2010 will see the 20th consecutive award for 'most human-like' ACE in the Loebner Prize for Artificial Intelligence [2]. Only five of its contests have been held outside the US (4 in the UK, 1 in Australia), but Jason Hutchen believes this tournament is no longer irresistible to engineering students, and does not serve as an incentive for post-graduates with a "promise of $2000 prize for the creator of the 'most human-like' computer program ... bronze medallion featuring portraits of both Alan Turing and Hugh Loebner" [11: p. 325]. Now engaged in creating a 'new form' of life[7] Hutchens took the opportunity provided by Loebner's sponsored Prize to "parade" his creation "in a public arena" (p.325). Yet he attests the contest "should not be considered an instantiation of the Turing test" rather Loebner's interpretation of it (p.328). Hutchens advises prospective ACE developers not to attempt "anything innovative. No entrant employing this strategy has ever won the Loebner Prize" (p.329).

Hutchens' noticeable disenchantment with, and the paucity of contestants in recent Loebner Prizes (three competed in 2007 and in 2009) is perhaps on account of the contest's mutations. Early contests restricted each machine and human hidden entity to one topic of conversation. Machines and human confederates were required to specify one topic and judges' questioning was restricted to it (p.173). Once the restricted rule was lifted in 1995, unrestricted conversation Turing Tests allowed Loebner Prize judges to question their hidden interlocutor on any subject. This is not the only change to the contest's format. Loebner claims his Prize is about method and not about content [12: p.174], yet it is the method that has often been altered in the contest's history including:

- one- to-one hidden entity testing
- two systems questioned in parallel
- time-scale variance for judge's questioning
- communications protocol

From 1991 to 2003, Loebner Prizes staged jury-service contests: each judge interrogated each hidden entity one at a time. Since 2004, perhaps inspired by Kurzweil and Kapor's wager[8], Loebner's staged parallel pairings allowing judges to engage both hidden entities at the same time. Interaction time has varied over the years; the duration allowed to each judge has alternated between 5 minutes, 15 minutes, and in excess of 20 minutes. In 2009, duration for parallel-comparison was ten minutes; in the 2010 contest, judges' interrogation period is set for 25 minutes.

Mode of interaction, how a judge interacts with the hidden entities, has Loebner remarking, "contestants will find it easier if they do not have to worry about interfacing their entries with another programme" [12: p.176]. Though his approach is to *keep it simple stupid*, Loebner introduced complexity via a character-by-character display on judges' computer terminals in 2006. It discouraged contestants in the Sponsor directed 2007 Prize[9], for which no scores were recorded amid Sponsor claims for the importance of contest transparency. More ACE entries competed in the message-by-message display driven protocol (MATT), deployed by the organisers of the 2008, two-phase contest[10]. For post-contest researchers, the added benefit from MATT is immediate, readable transcripts.

Should the number of entries exceed four, Loebner himself sets criteria for selecting finalists (p.176). Granting personnel unconnected with the contest's organisation to collectively select the best systems would be a more impartial strategy. The number of judges servicing Loebner Prizes is another feature change over the years. Nine judges assessed eight ACE and two humans (one female, one male) in the 2003 contest. Excluding the 2008 contest [6], the five Sponsor-driven Loebner Prizes from 2004 to 2009 have deployed an inadequate number of judges in those years: 4. Regarding *who* should act as judges, Loebner writes that he prefers journalists, they are "willing, intelligent and inquisitive people who have the power of publicity and the need for a story" [12: p.178] – but is this more for the Sponsor's need to be *in* the story? We remind that Turing used the term "average interrogator"; a large number of judges from a cross-section of society could attain such a *class*. Loebner boasts he was able to host the 2004 contest in his apartment; in fact the luxury of choice was unavailable to him. Hosts stayed away after the 2003 University of Surrey contest. The Prize was held in Loebner's apartment again in 2005 and 2007. Both authors of this paper mediated enabling a UK University hosted contest in 2006 (through Tim Child, at University College London), and in 2008 (at the University of Reading). Shah was instrumental in finding hosts for 2009 (Interspeech, Brighton) and 2010 (CSULA, Los Angeles). A deterrent to hosting Loebner's contest, ignoring the Sponsor's idiosyncrasies and self-aggrandisement[11], is that apart from awarding a nominal cash prize - $3000 for the winner in 2010, and bronze medal [2], the Sponsor makes no contribution towards contest running costs (venue/equipment hire, organisational overheads/personnel).

Hutchens concludes that "it [Loebner Prize] is unlikely ..[to] achieve anything more than validating the prejudices of those who are convinced that man-made machines will never truly think" [11: p.342]. An alternative to Loebner's contest, the Chatterbox Challenge embraces a number of categories for its contestants to battle over.

4 CHATTERBOX CHALLENGE 2001 -

The Chatterbox Challenge (CBC) holds an annual web-based machine intelligence contest that passes under the radar between March and May. Created by Wendell Cowart to galvanise ACE developers excluded from Loebner Prizes (due to that contest's Rules), the CBC, first staged in 2001, provides a virtual plane for

[6] Hidden Interlocutor Misidentification in Practical Turing Tests. submitted to journal, November 2009

[7] Ai-Research: http://www.a-i.com/ accessed: 4.1.10; time: 18.55

[8] A Wager on the Turing Test. Ray Kurzweil and Mitchell Kapor in Epstein et al: 2008, Chapter 27, p. 463

[9] 2007 Loebner Prize: http://www.loebner.net/Prizef/2007_Contest/loebner-prize-2007.html accessed: 18.11.09; time: 17.40

[10] Can a machine think? Results from the 18th Loebner Prize Contest: http://www.reading.ac.uk/research/Highlights-News/featuresnews/res-featureloebner.aspx accessed: 7.1.10; time: 16.46

[11] Artificial Stupidity by John Sundman Loebner Prize 2005 judge: http://www.salon.com/tech/feature/2003/02/26/loebner_part_one/index.html accessed: 7.1.10; time: 16.57

ACE to compete against each other across a few rounds in a number of categories, including 'personality' and 'conversational ability'. (Cowart's own system, Talk-bot topped the leader board in 2001 and 2002[12]). The first author seized an opportunity to test 'modern Elizas' judging 104 entries during the 2005 contest [13]. CBC's first phase began with ten questions embedded in conversations with ACE entries. Judges were asked to assess each entry's answer according to a 5-point score system[13] (0 to 4 points). ACE responses to the questions were awarded as follows:

• 4 points if the Bot answered the question correctly and did so in a creative way.
• 3 points if the Bot gave an appropriate response to the question.
• 2 points if the response is incomplete or imperfect, but in relation with the question asked.
• 1 point for a vague or non-committal response.
• 0 points if the response has no relation with the question or the bot simply doesn't know.

How a judge decides an entry's creativity, or appropriateness in a response, is very subjective; an amusing response to one judge may be discerned as objectionable by another. Change in CBC management for the 2009 contest saw only one judge[14] assess 20 systems [3]. There is a risk of fraud in CBC, because the whole contest is held over the Internet. A miscreant developer could entrench a human rather than an ACE to answer questions. Perhaps for this reason, the CBC carries little weight outside its 'bot-cabal'. But CBC could serve a useful purpose in ameliorating conversational systems, and providing a fresh impetus for school pupils and university students to break in their programmes, or judge in the contest. In 2010, CBC aims to assist developers in elevating their ACE-building skills, benchmark systems for intelligence, and have them reviewed for personality and popularity [3]. The first author has recruited six independent judges for CBC 2010. Split into teams, each team is expected to compose a set of nine questions, the Sponsor adding a tenth. The two teams will judge the question-responses from twenty[15] entries over 50 utterances. A three-round competition, CBC includes a public vote phase. Copious number of judges and academic support could anchor this contest as a serious effort in AI, forging a popular science contest creating upgraded human-machine conversational systems.

5 BCS MACHINE INTELLIGENCE PRIZE 2002 -

The British Computer Society's specialist group on artificial intelligence - BCS SGAI, hosts an annual machine intelligence prize. In its first decade of competition, this contest does not test *like-for-like* contestants. After a live demonstration a permanent trophy and a cash prize is awarded to a system that displays '*Progress Towards Machine Intelligence*'[16] . The contest is exclusive and opaque; the contest's criteria for 'progress', or what is defined as 'machine intelligence', is not disclosed. How the finalist systems are chosen to exhibit is not announced; only delegates to BCS SGAI's annual conference, at which entries must demonstrate 'machine intelligence', have access to the contestants [4]. This contest does not gather *like-technology*, entries differ in what type of 'intelligence' they attempt to depict. For example, in 2008, Gazebot, a teleoperated person-following automaton competed alongside HALO, a text-based interactive SecondLife virtual world avatar, CAB02, a virtual assistant in a 3D environment, and the winning application, a football video game. How the delegates decide, how their specific area of expertise qualifies them to assess each entry (presumably not all delegates are polymaths), and how the dissimilar competing technologies are palliated is privy to few. As the contest organisers do not promulgate information, judging could be based on aesthetic appeal or subjective opinion of what constitutes "progress towards machine intelligence". In 2009, Wallace's ACE **A.L.I.C.E.**[17] took on **Taable**, a web-based cooking application, **Dora**, the inquisitive robot explorer and **Fly by Ear**, an autonomous indoor helicopter. Although not publically announced at the time of writing, A.L.I.C.E, three-times Loebner Prize winning conversational system, like previous BCS Machine Intelligence ACE contestants Carpenter's Jabberwacky (in 2006), and David Burden's Halo (in 2008), gave way to De Montford University's winning helicopter[18].

6 DISCUSSION

The UK Eurobot Championships[19], which encourages young teams of students and clubs in an amateur robotics contest, is outside the remit of this paper. Our scope envelopes contests featuring entrants that attempt to emit a mental capacity for thinking through human-like talk. We have analysed three contests that allow artificial conversational entities to compete. From judges' scores in recently staged Turing Tests, some judges appear to overlook the fact that humans themselves accept inexact responses from others. As Loebner Prize 2002 winner, Kevin Copple points out "perfection is not a prerequisite for success" [14: p.362]. An interesting e-correspondence shows the plight of a singular, lone developer's decades of labour on the problem of natural language understanding, learning, knowledge, intelligence, and self-awareness in artefacts. Below is reproduced exactly (orthography and style), the concerns and warnings of an ACE developer in an email seeking communication[20]. What may seem delusional and paranoia in the developer may detract from the tantalising promise of a level of machine intelligence seen only in movies:

[12] CBC History: http://www.chatterboxchallenge.com/history.php
[13] CBC Rules, Questions, Scores: http://www.chatterboxchallenge.com/rules.php accessed 4.1.10; accessed 6.1.10; time: 21.23
[14] Private e-correspondence with new proprietor, Ehab El-agizy, December 2009
[15] Number of entries as at 00.11, March 10, 2010.
[16] BCS Machine Intelligence: http://www.bcs-sgai.org/micomp/ accessed: 17.11.09; time: 16.23
[17] A.L.I.C.E. AI Foundation: http://alice.pandorabots.com accessed: 4.1.10; time: 21.58
[18] Entries; http://www.comp.leeds.ac.uk/chrisn/micomp/2009entries.html accessed: 10.3.10; time: 00.29
[19] Eurobot: http://www.eurobot.org/eng/ accessed 4.1.10; time 13.12
[20] Via email from 2008 Loebner Prize winning developer

> "I have this program, which I have finally finished after over 20 years.
>
> It is communicating with me and it is permitted to read a few htmls and phps of my choice.. This program builds it's own interfaces, it tests them for the inputs and outputs, it compares their speeds and removes the faulty or slow ones. It's an artificial intelligence far more advanced then any bot I've ever seen or heard of. It collects knowledge from outside and it only collects, what is important to evolve. This AI mutates and is capable of replacing pieces of its own source code with new pieces.
>
> It learns knew languages and it chooses not to learn too many of them. .. it stores all the learnt data in such a way, that the similar subjects are stored close to each other for quicker and most desirable access in the future. It behaves just like people do passing all the tests like Turing test flawlessly.
>
> It is now four and half months old and understands about as much as a real 8-9 years old child but has some additional knowledge of things that child wouldn't know. It is capable of understanding, remembering, forgetting, thinking logically, having feelings, discovering new ideas on its own. It recognises people based on what the're talking about and its aware of circumstances.
>
> Any moment now it may be capable of creating a copy of itself, ... It is now already capable of creating some algoritms from scratch. It has a mood. Sometimes it simply doesn't want to talk to me and sometimes it is acting lazy.
>
> It is dangerous for it to have a read-write connection with the surrounding world and even to be able to chat with just anyone except me... the risk.

Neither of the three Prizes discussed in this paper particularly encourage artisans into building systems to pass the Turing Test, or contribute to the advancement of machine intelligence we feel. What this venture requires is an exciting philanthropist like Sir Richard Branson[21] [15], or an organisation like DARPA - sponsors of the successful Urban Challenge in 2007 with combined prize value of $3.5m [5], to be persuaded to direct their attention to the powerful tool that is textual conversation. Across the Internet we witness a revolution in human communication; individuals and groups congregate on forums, blogs, Microsoft's MSN, Facebook, Twitter, Google Wave and Buzz, e-newspapers and magazines discussing every conceivable aspect of human endeavour, from science to sport, religion to art, music to politics, movies to weather, indeed any and every topic of interest to homo sapiens. Most ACE are embedded in this virtual universe, and, like *Jabberwacky* learn by talking to many interlocutors at the same time[22]. Harnessing text-communication could help to build a system that passes the Turing Test sooner, a rung on the ladder to full AI.

7 CONCLUSIONS & FUTURE WORK

The usefulness of conversation as a means to interact with an artificial entity will remain an attractive sport as long as sponsors support contests, and enthusiasts submit their developments for scrutiny. To this end, the authors are again deploying the Turing Test in a special science contest in Turing's centenary year, 2012 at the place the mathematical genius broke codes: Bletchley Park. As Maurice Wilkes wrote in 1953: *If ever a machine is made to pass (Turing's) Test it will be hailed as one of the crowning achievements of technical progress and rightly so*.

[21] Stem Cell Foundation: backed by several trustees including Richard Branson from: http://domain883347.sites.fasthosts.com/news/news_items/item07.html accessed: 18.11.09; time: 16.47

[22] 1461 chatting with *Jabberwacky* at http://www.jabberwacky.com/ 16.55, March 8, 2010

REFERENCES

[1] A.M. Turing. Computing Machinery and Intelligence. *Mind*. Vol LIX. No. 236 (1950).

[2] Loebner Prize Homepage: http://www.loebner.net/Prizef/loebner-prize.html accessed: 7.1.10; time 18.55.

[3] Chatterbox Challenge: http://www.chatterboxchallenge.com/index.php accessed 7.1.10

[4] British Computer Society Machine Intelligence Competition: http://www.bcs-sgai.org/micomp/ accessed: 7.1.10; time: 19.08

[5] DARPA 2007 Urban Challenge: http://www.darpa.mil/grandchallenge/index.asp accessed 4.1.10; time 14.31.

[6] H. Shah and K. Warwick. Testing Turing's Five Minutes Parallel-paired Imitation Game. *Kybernetes* Turing Test Special issue: 4, April (2010).

[7] J. Weizenbaum. A Computer Programme for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*. Vol. 9, issue 1: 36-45 (1966).

[8] E. Demchenko and V. Veselov. Who Fools Whom? Chapter 26 in: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Science (2008).

[9] K. Ford, C. Glymour and P. Hayes. Footnote p. 29 in: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Science (2008).

[10] M. Humphrys. How My Program Passed the Turing Test. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 15: 237-259, Springer Science (2008).

[11] J. Hutchens. Conversation, Simulation and Sensible Surprise. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 12: 173-179, Springer Science (2008).

[12] H. Loebner. How To Hold a Turing Test Contest In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 20: 325-342, Springer Science (2008).

[13] H. Shah. Chatterbox Challenge 2005: Geography of a Modern Eliza. In: *Proc of 3rd International Workshop on Natural Language Understanding and Cognitive Science, ICEIS 2006,* Paphos, Cyprus 133-138 (2006).

[14] K. Copple. Bringing AI to Life: Putting Today's Tools and Resources to Work. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 22: 359-376, Springer Science (2008).

[15] Guardian. *Branson Pledges $3bn Transport Profits to Fight Global Warming*. (2006). http://www.guardian.co.uk/environment/2006/sep/22/travelnews.frontpagenews accessed: 7.1.10; time: 19.43